

**PARIS SCHOOL OF ECONOMICS  
MASTER IN PUBLIC POLICIES AND DEVELOPMENT**

**ECONOMETRIC TEAMWORK**

**TOPIC:**

**THE IMPORTANCE OF TRACKING IN LONG-TERM HOUSEHOLD PANEL SURVEY: EVIDENCE FROM THE  
IMPACT OF ORPHAN-HOOD ON HUMAN DEVELOPMENT IN RURAL TANZANIA**

**WRITTEN BY:**

**GEORGES VIVIEN HOUNGBONON (h.gvivien@yahoo.com)**

**SEBASTIAN GUENDEL ROJAS (sguendel24@hotmail.com)**

**VIET ANH TRAN (vietanhtran.ftu@gmail.com)**

# Table of Contents

- Introduction..... 1
- I . When panel surveys care about tracking: following-up the human capital of orphans and non-orphans ..... 2
  - A. The orphan hood crisis: some background information..... 2
  - B. The Kagera Health and Development Survey: efforts to obtain proper panel data..... 3
  - C. The importance of tracking: Attrition in our sample ..... 3
- II . Dealing with attrition in longitudinal panel surveys ..... 5
  - A. When is attrition a problem: the distinction between two cases ..... 5
- Case 1: Non-random attrition is a problem of selection on observables ..... 6
- Case 2: Non-random attrition is a problem of selection on unobservables ..... 6
  - B. Empirical strategy: how to assess the problem of selection in the KHDS survey and how to deal with it..... 6
- a. Assessing whether attrition is a problem or not for our estimations ..... 7
- b. Assessing the source of non-selection ..... 8
- c. Assessing alternatives to solve the problem of selection on unobservables due to attrition .. 8
- III . Results: The impact of tracking individuals on the estimable causal impact of orphan hood.. 9
  - A. Preliminary assessment of the impact of tracking on the estimates..... 9
    - a. A descriptive attempt to detect selection..... 9
    - b. Making sure that we are comparing similar samples: running descriptive statistics ..... 11
  - B. Econometric estimation: Assessing selection and importance of tracking ..... 14
    - a. Improving our first attempt results to detect selection..... 14
    - b. Exploring tools to correct for selection bias..... 15
    - c. Unveiling the source of selection ..... 22
- Conclusion ..... 24
- APPENDIX: ..... 26
- REFERENCES: ..... 27

## Introduction

In sub-Saharan African countries, greatly affected by HIV/AIDS, children orphan hood has been perceived as a major source of poverty, as the negative impacts of orphan hood in children's human development risk to perpetuate into adult life. Determining and understanding whether there is or not a causal link between orphan hood and reductions in children's human development is thus crucial to set the policy alternatives that will fight the impact of this condition on future poverty. If it happens to be a link between these two phenomena in countries with high mortality rates then global and national initiatives that would omit this factor from their strategies would certainly miss a crucial part of the story, thus providing misleading alternatives to fight poverty.

Empirical determination of this link has not been easy however. This has been mainly related to the quality of the data used, and in particular, the problem of attrition in panel surveys.

Initially, the use of cross-sectional data was identified as the first source of difficulties. Such type of data can indeed lead to erroneous conclusions if pre-orphan characteristics have also an impact on the outcome of interest and at the same time on the orphan hood status, thus leading to an omitted variable bias. Cross-sectional analyses might also lead to other omitted variable biases if orphans are strategically placed in better-off households. Then, any differential in outcomes between orphans and non-orphans would have to take into account differentials in outcomes due to new living conditions as a result of parental death. Adding household fixed effects under such a framework is not a very good idea either since strategic placement of orphans might lead to comparisons between orphans and non-orphans within the household which is not completely random.

The literature tackled these issues by using panel surveys instead of cross-sectional data since such type of data is appropriate to deal with pre-orphan omitted variables. For example, Evans and Miguel (2007) used panel data from Kenya to study a large sample of non-orphans enrolled in grades 1-7 in 1998 and re-interviewed in 2002. They assessed the impact of orphan hood transitions on schooling participation. They showed that maternal deaths led to lower participation, as well as in the 1-2 years before death. On the contrary, paternal orphans did not have lower school participation. Authors like Kathleen Beegle, Joachim De Weerd and Stefan Dercon (2008a, 2008b, 200c) have been also highly proactive in the use of this type of data showing that orphan hood matters in the long run for health and education outcomes.

But panel data has not been the panacea. This has been mainly due to the problem of attrition that may provide misleading results because of problems of selection. The literature has also tried however to tackle this issue by investing in panel surveys that follow-up individuals in case of migration. This has been the case of the survey used in this paper, the Kagera Health and Development Survey implemented by the World Bank in Tanzania. Thanks to this survey, researchers analyzing the issue of orphan hood have been able to deal straightforward with attrition in their estimations. For example, Beegle, De Weerd and Dercon (2008) using a sample of 718 non-orphaned children surveyed in 1991-1994, who were retraced and reinterviewed as adults in 2004, found that maternal orphan hood had a permanent adverse impact of 2 centimeters of final height attainment and one year of education attainment.

Although these advances in panel surveys have been crucial to the identification of the impact of orphan hood on human development indicators it remains yet unclear the magnitude of the impact of tracking individuals on the assessment of orphan-hood effects. Until now no research paper has provided an answer to this question which in our opinion is crucial to justify investments in such types of surveys. It might be probably the case that the impact of tracking on the estimates might not be that important and thus spending a lot of money in these surveys might result in superfluous investments.

On the contrary, it might be the case that such investments are necessary because attrition might cause huge problems of selection bias. If this is the case then such investments are crucial to obtain good causal implications of the impact of orphan hood on human development indicators. Having this information will let us certainly fight better sources of future poverty, such as orphan hood.

In light of these introductory observations, the goal of our paper is twofold. In the first place, we will try to answer the question that has remained unanswered until now: what is the impact of tracking on long-term panel surveys. We will do this by studying the specific cases of education and height as proxies for human development of the population under scrutiny. If it is the case that tracking is important (we are going to see that it is the case) then, in the second place, we will provide information about how to deal with selection due to attrition. This data is indeed ideal to answer this second issue since due to tracking we are able to observe what would have remained unobserved because of attrition. We will see that standard methodologies to deal with attrition, such as the Heckman methodology, would have led us to misleading results.

The structure of our paper is the following: in the first part we describe the structure of our data, and in particular, the way tracking was set up. In the second part, we explain in which conditions attrition is a problem. In the third part, we present our empirical strategy to answer our two questions. The fourth part presents our results and concludes.

## **I. When panel surveys care about tracking: following-up the human capital of orphans and non-orphans**

### **A. The orphan hood crisis: some background information**

The issue of orphan hood has become extremely important in Tanzania. According to the Tanzania Reproductive and Child Health Survey (TRCHS), among the 15 million children under the age of 15 in 1999 the number of orphans was estimated to be 1.65 million, of which 165.000 were double-orphans, 960.000 were parental orphans, and 525.000 were maternal orphans. In terms of proportions, these figures are astonishing. The fathers of 6.4% and the mothers of 3.5% of children under 15 had died, of which nearly 1.1% lost both parents.

Although, these figures include AIDS and non-AIDS orphans, HIV/AIDS has been seen as the major contributory factor. The increase in adult mortality due to HIV/AIDS has strengthened significantly during the 1990s. As a matter of fact, according to the MEASURE project held by the National AIDS Control Programme of Tanzania and the National Bureau of Statistics, and other studies, HIV/AIDS became in the late 1990s the leading cause of death among adults in Tanzania, driving thus the rise in the orphan hood crisis.

The need to understand the impacts of such crisis has become pressing, and many studies and projects have tried to respond to this challenge.

The main efforts have come from the Kagera region, the first one to be seriously hit by this epidemic in the late 1980s. The Kagera region announced indeed the first case of HIV/AIDS in Tanzania in 1983, and as of 1987, it was one of the first regions to report high prevalence of this disease. Although, these high prevalence rates have diminished during the 1990s, the region has made enormous efforts to recover statistics about the way this tragedy evolved during this period, making this information crucial to understand phenomena such as the orphan hood crisis. As suggested by the Tanzania Commission for AIDS and the National Bureau of Statistics, orphan hood rates in this region have been comparable to the national average; 10.5% versus 10.8% respectively, making this information even more crucial to understand this specific phenomenon.

## **B. The Kagera Health and Development Survey: efforts to obtain proper panel data**

The Kagera region powered by the World Bank has invested in particular in longitudinal panel surveys, such as the Kagera Health and Development Survey (KHDS). The Kagera Health and Development Survey consists of two round of surveys (1991/1994 and 2004), the KHDS-1 and the KHDS-2, that contain information on many issues going from anthropometric measurements, to price data on local markets, to consumption, migration, and others. Moreover for all KHDS-1 respondents who died between 1991/1994 and 2004 information on the circumstances of their death was collected. In principle this data should make it easier for researchers to obtain causal implications about how AIDS, as well as other conditions, could have affected individuals in this region since we can account for pre-temporal characteristics.

Although panel surveys exist in other countries and regions, the novel aspect of this data is that it was conducted taking care of important issues such as attrition, by following-up all the individuals possible of the first round into the second round.

According to KHDS team, the purpose was to find out as much information as possible on the 2004 location of the original respondents. Out of the 919 original households, information was lost only for 42 of them. For the remaining ones, households were contacted or sufficient information obtained to trace them where ever they currently resided. It was found in 2004 that new households were created by some of the original households' members. These household members, who were dependent children 10 years ago now live independently; others got divorced, orphaned or went to live with extended family. At end-line there were 3051 households. Only 45% of these households were found in the original villages, 30% had moved within the region, 14% to another region in Tanzania and 2% had moved to another country. Within other regions of Tanzania most households were found living in Mwanza, followed by Dar-es-Salaam, Shinyanga and Mara. Those who had moved to other countries were mostly found living in Uganda, Rwanda and Kenya, but some were reported to be living as far away as Norway, Germany, Sweden and the United Kingdom. It was found that about 14% of the total population had died. Therefore, had households not been tracked, the rate of attrition would be 55%.

Although is not yet clear whether tracking is important or not to have better causal estimates, as we are going to try to find out later on, following-up individuals has been a great asset to obtain good quality panel data. In comparison with most longitudinal surveys, it seems that these efforts were worthy: in 93% of baseline households, at least one household member was re-interviewed in 2004. Among all surviving members of the baseline households, about 82% were re-interviewed in 2004.

These efforts to follow-up individuals seem important because migration and dissolution of households are often specific processes that can induce a bias in the final estimates. Thus, we expect that by going through this costly exercise we would be able to deal with these sources of bias.

## **C. The importance of tracking: Attrition in our sample**

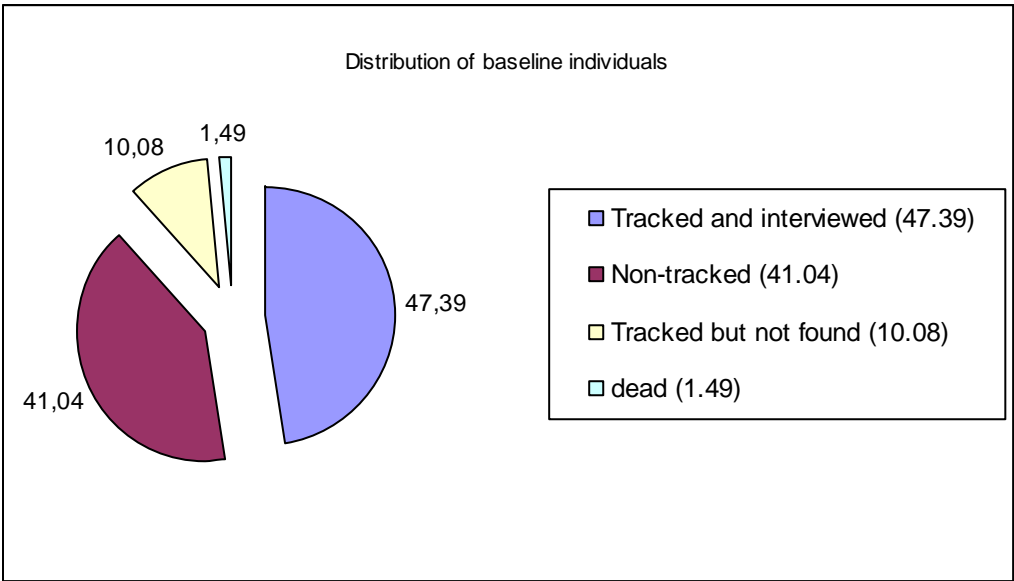
If we move now to our sample of non-orphans, we can see that the gains of having followed-up individuals in terms of the quality of the data are huge.

Our longitudinal panel of orphans is constituted of two points in time, 1991 and 2004. Since we are analyzing the impact of having become orphan during this period (1991-2004), our baseline sample in 1991, contains only non-orphans. Because some of these children lost their parents, our end-line in 2004, contains orphans and non-orphans all of them under the age of 28 (including the 28 year-olds).

As we said previously, attrition is a fact in longitudinal panel surveys, and our panel is not the exception. Attrition in our sample was caused for three reasons: death, migration, and missing people.

Figure 1 shows that the impact of having followed-up individuals on the quality of our data is simply astonishing. While we miss information in 2004 for some of the 1991 non-orphans, either because they were missing (10.08%) or because they died (1.49%), thanks to the efforts of tracking we were able to obtain information in 2004 for 47.39% of the 1991 non-orphans that otherwise would have been missing. This means that had we not tracked non-orphans, we would have missed information for nearly half of the sample of 1991. As said before, this is impressive by the standards of longitudinal panel surveys since these panel surveys often present high levels of attrition.

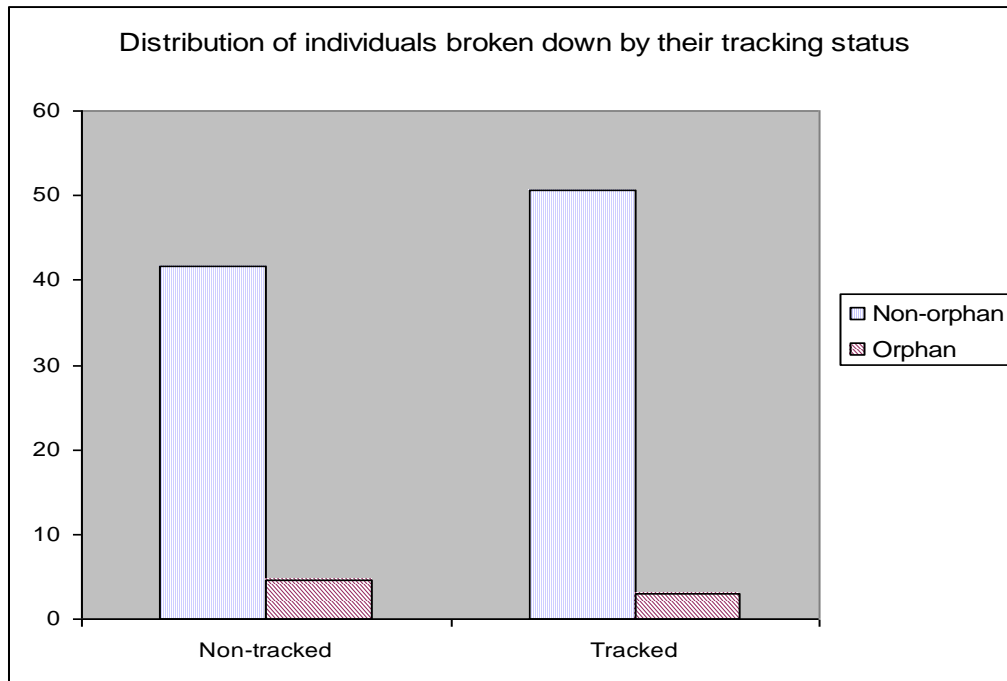
**Figure 1.** The Impact of Tracking Individuals on the Quality of our Panel Data



By looking at figure 3, if we break down our sample of orphans and non-orphans in 2004 by their tracking status; the impact of having followed-up individuals on the quality of the data is even more revealing. For example, had we not tracked non-orphans between 1991 and 2004, of the 92.43% of non-orphans in the sample, we would have missed information for a little bit more than 50% in 2004. Likewise, had we not tracked orphans during this period, of the 7.57% of orphans in the sample, we would have missed information for a little bit more than 38% of this sub-sample of individuals.

This means that because of the efforts to follow-up individuals we are able to know that attrition is more important among non-orphans than among orphans. Although we cannot know yet whether attrition could entail a bias on our estimates, having this information before-hand is important because it tells us that probably attrition in our sample follows a non-random process, thus being a potential source of bias for our estimations.

**Figure 3.** Orphan hood Status broken down by Tracking Status



The main conclusion that we can make about these patterns of attrition is that tracking individuals is very important to have high-quality data. Thanks to this, we are able to reduce missing observations, thus making our panel more balanced. The question is: is this good news for any estimation that would use this data? This is what we are going to try to find out in the coming sections: is attrition in our sample a problem? What can we do about it if this is the case?

## II. Dealing with attrition in longitudinal panel surveys

### A. When is attrition a problem: the distinction between two cases

Although having lost a proportion of the sample as a result of attrition might be a problem for many reasons, there is no necessary relationship between the proportion of this sample loss and the existence or magnitude of a bias due to attrition. As a matter of fact, even a large amount of attrition causes no bias if the reasons people left the sample are random. But this might be not the case, and thus, when non-random attrition happens we can have problems of selection or self-selection bias. More precisely we are going to distinguish two cases to make the argument tractable: selection on observables and selection on unobservables.

We can take the standard textbook parametric model, to describe these processes.

$$Y_{i,t1} = \alpha + \beta \text{orph} + X_{it0}\gamma + u_i \quad (1)$$

$$A_i^* = Z_i\delta + v_i \quad (2)$$

$$A = 1 \quad \text{if } A^* > 0 \quad (\text{if individual has been tracked})$$

$$A = 0 \quad \text{if } A^* \leq 0 \quad (\text{if individual stayed})$$

$Y_{i,t1}$  is the outcome in 2004,  $orph$  is equal to 1 if the individual is orphan and 0 otherwise.  $X_{i,t0}$  is the vector of controls at baseline (1991) and  $u_i$  represents the residuals of the equation. Equation (2) represents the relation between the latent variable  $A^*$  and the determinants of migration  $Z_i$ . In fact,  $A$  is the dummy variable of selection into the sample indicating whether an individual has stayed or gone. We assume that the observables in equations (1) and (2) are jointly independent from the residuals  $u_i$  and  $v_i$ . This assumption includes independence between the observables of equation (2) and the unobservables of equation (1). In case this latter is not true, we can always control for the observables at stake in equation (1).

$Y_{i,t1}$  is observed if  $A=0$ , meaning that we assume that we only observe the outcome of those who stayed. Our goal is to assess whether or not we can obtain a non-biased estimate of the impact  $\beta$  of orphan-hood on the outcome. The expectation function of the equation of interest using the available observations reads:

$$E(Y_{i,t1} / orph_i, X_{i,t0}, A_i = 0) = \alpha + \beta orph_i + X_{i,t0} \gamma + E[u_i / v_i \leq -\delta Z_i(orph)] \quad (3)$$

Expression (3) raises two cases in which selection due to attrition would bias our estimate of  $\beta$ .

#### **Case 1: Non-random attrition is a problem of selection on observables**

If the observables in the selection equation are correlated with the observables in the equation of interest; then selection is on observables and the expectation function is not correctly estimated. For example if  $Z_i$  depends on orphan-hood status or if the determinants of migration are correlated with orphan-hood status then the conditional expectation of the residuals in the equation of interest is different from zero and then the impact of orphan-hood is biased. However, this can be easily solved. As said before, the problem of non-random selection on observables can be solved by adding the scalar  $Z_i$  in the equation of interest, following the ignorability assumption. So, as long as we can control for the observables to cause selection, attrition would not cause any problem, and this will be as good as random.

#### **Case 2: Non-random attrition is a problem of selection on unobservables**

If we cannot control for observed characteristics to correct for selection due to attrition, this means that our estimation of  $\beta$  is still biased. More concretely, selection bias may be due to a correlation between the residuals of both equations:  $Corr(u_i, v_i) \neq 0$ , so even if  $Z_i$  is independent from the observables in (1), we still have a bias. This is a selection on unobservables and it can be solved by estimating the bias due to selection following a complete parametric model, such as in Heckman.

### **B. Empirical strategy: how to assess the problem of selection in the KHDS survey and how to deal with it**

The goal of our analysis with regard to attrition is threefold. First of all, we assess the extent to which attrition is a problem or not in our sample, i.e. we look at whether attrition follows a non-random process in our sample. Second, in light of our previous discussion, we try to assess the source of non-random attrition. Finally, we discuss the alternatives that we have to solve for attrition bias in our sample. We then try to obtain more general implications on this.

By looking at our previous graphical analysis we suspect that attrition could be non-random in our sample, however we need to study this with more attention. This section will try to set the strategy that we will follow to do this.

In the coming sections, the analysis of the impact of attrition (tracking) on the estimates will be done using two relationships: the impact of orphan hood on height on one hand, and its impact on education on the other.

**a. Assessing whether attrition is a problem or not for our estimations**

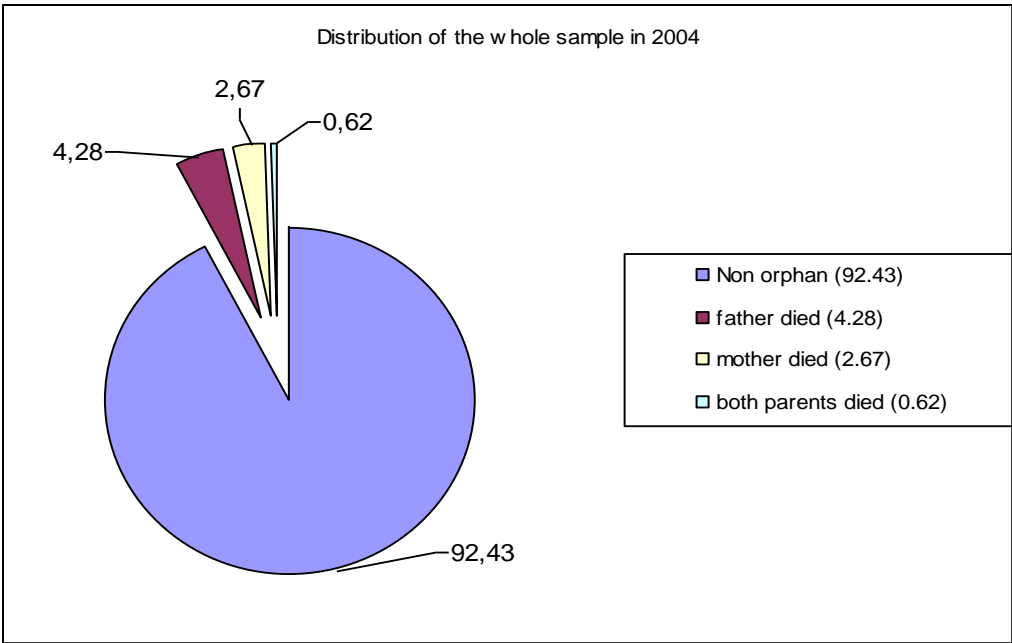
We want to assess first whether attrition is or not a problem for the empirical assessment of these relationships in our sample. Checking for this has been usually difficult for the simple reason that attrited individuals do not usually provide information. In our case, we are not penalized by this handicap since thanks to the specific structure of our data we are able to have information on these individuals, as if we had not had any attrition at all.

As we mentioned previously our data provides information on those individuals who were tracked. Because we are able to identify these individuals, we are thus able to compare two types of situations: the standard situation in longitudinal panel surveys in which we are missing nearly half of the individuals surveyed in the first round of the survey, and the other, more unusual, situation in which we would have the entire population surveyed in 1991.

We test for the impact of attrition in a simple and descriptive way. We do this first by comparing the means of height and education in both samples (the sample with tracked and non-tracked individuals; and the sample with only non-tracked individuals) to see if they differ. If they differ, then we have basic evidence that attrition might be non-random in our sample.

We shall make clear at this point that we make these comparisons using three types of populations as we can see in figure 4, since we want to know whether the impacts change according to the type of orphan hood status. We thus compare in both samples, the one including the whole population (tracked and non-tracked individuals) and the one including only the non-tracked individuals, the average height and education of orphans and non-orphans (any parent died), but also that of parental orphans against that of non-orphans, and that of maternal orphans against that of non-orphans.

**Figure 4.** Structure of our two samples



This step in our analysis is crucial since, as mentioned before, if the results are similar in both samples then this would give us some evidence suggesting that attrition might not be a problem in our case, and thus that this event follows a random process in our sample.

#### **b. Assessing the source of non-selection**

However, if there are differences in mean height or mean schooling between our two samples, it might be the case that these differences that we observe are not due entirely to attrition. It might be probably the case that orphans and non-orphans within both samples are initially different at baseline in 1991 (in terms of age for example), so the differences in mean height and mean schooling in 2004 that we observe are due to these initial discrepancies between individuals and not to attrition.

In order to check that the differences we observe between our two samples are not due to these initial characteristics, we run descriptive statistics on some baseline characteristics of orphans and non-orphans within each sample. If we see some differences between orphans and non-orphans with regard to these baseline characteristics, this will probably suggest that the differences we observe between our samples are due to discrepancies between individuals concerning these baseline features, and not to attrition per se.

If we condition our conditional expectations on a vector of variables  $Z_i$ , then we hope, this will help us control for selection on observables in our estimation, and thus obtain similar estimates of mean height and mean schooling in both samples. If we think about attrition as some kind of treatment, as in a Rubin Causal framework, then we are looking at whether treatment is ignorable conditional on these observed characteristics.

More concretely, we run OLS regressions controlling for these characteristics in which orphans and non-orphans differ. If after running these regressions we still observe differences between the two samples, whatever the orphan hood status, then this is evidence that our selection due to attrition is on unobservables, and this is causing anyway a bias. So we need to find a way to correct for this bias.

#### **c. Assessing alternatives to solve the problem of selection on unobservables due to attrition**

If selection is on unobservables, this means that we can proceed within the framework of a parametric approach.

Now, following this approach in our case is very interesting for one single reason. Since we know the whole sampled population of 1991 non-orphans in 2004, this means that we are able to know the true reasons of attrition. Indeed, as we are going to present in the coming sections this in the data. The goal of this final step of our analysis is thus to check whether a full parametric approach is indeed the correct way to correct for the attrition bias.

Actually, the Heckman procedure could be a useful way to correct for selection bias, if any. However, this procedure assumes that we have complete information at the baseline survey that can help infer the migration decision of individuals. If we could obtain this information, then we can correct for selection bias in case individuals would have not been tracked. As a consequence, whether tracking individuals would be crucial in our analysis or not, this would depend on the accuracy of these other alternatives to correct for the selection bias due to attrition. These alternatives could be a Heckit method - if selection is on unobservables - or sample reweighting - if selection is on observables. Although these solutions might totally work out for some specific topics and then conclude that tracking is not important, it is not true however that the same would apply to all related topics in

question, suggesting thus that tracking might be useful for these other issues at stake. We focus in our study on how tracking may affect the impact of orphan-hood on human capital and discuss whether it would be possible to correct for selection bias in order to conclude on the importance of tracking in a household panel survey.

### **III . Results: The impact of tracking individuals on the estimable causal impact of orphan hood**

We first need to compare the impact in the whole sample with the impact in the sample with non-tracked individuals. This will help us conclude whether there is a selection bias affecting the estimable impact of orphan hood in the sample of non-tracked. In case there is a selection bias, then the second step will be to see if it could be corrected had individuals were not tracked. Hence we can conclude about the importance of tracking depending on whether there is a correction tool or not. Beginning with the first step, we are going to use simple descriptive statistics in order to have an insight about the discrepancy between the two samples.

#### **A. Preliminary assessment of the impact of tracking on the estimates**

##### **a. A descriptive attempt to detect selection**

This part presents the results from simple mean comparisons of height and schooling in 2004.

Let us describe before more precisely our data. Regarding the number of observations we could not find exactly the same figures as in Beegle, De Weerd, and Dercon (2008c). Contrary to the 718 children for whom they have complete socioeconomic and anthropometric data from baseline and follow-up, we manage to identify only 649 children for whom we have this information. In fact, we had 757 non-orphans within 6 and 15 years old at the baseline survey in 1991. Among these children only 92 could not be found and 16 died within the period from 1991 to 2004. Nevertheless, we still have 649 who were re-interviewed in 2004 and for whom complete data is available. Of these 649 non-orphans at baseline, 60% stayed in their previous location whereas the remaining 40% moved and were tracked in 2004. In addition, 118 out of the 649 lost at least one of their parents before reaching 15 years old. This represents 18.2% of orphans in the whole sample.

Table1 reports comparisons of average height and years of schooling of orphan and non-orphans both in the whole sample and in the sub-sample of non-tracked individuals. The last column of the tables is the most important. In fact, it reports the difference between the impact in the whole sample and the one estimated in the sample with non-tracked

Table 1. Average adult height and schooling of orphans and non-orphans

	Whole sample			Sample without tracking			
	Mean final height non-orphans (cms)	Mean final height orphans (cms)	Difference in mean final height (cms)	Mean final height non-orphans (cms)	Mean final height orphans (cms)	Difference in mean final height (cms)	Diff of diffs <sup>1</sup>
Any parent died <sup>2</sup>	160.5 (531)	157.8 (118)	2.7	163.4 (324)	154.1 (69)	9.3***	6.6
Father died	160.5 (531)	154.3 (68)	6.2**	163.4 (324)	153.6 (48)	9.8***	3.6
Mother died	160.5 (531)	163.9 (33)	-3.4	163.4 (324)	162.2 (9)	1.2	4.6
	Mean final schooling non-orphans (years)	Mean final schooling orphans (years)	Difference in mean final schooling (years)	Mean final schooling non-orphans (years)	Mean final schooling orphans (years)	Difference in mean final schooling (years)	Diff of diffs
Any parent died	6 (531)	5.5 (118)	0.5	6.6 (324)	4.7 (69)	1.9	1.4
Father died	6 (531)	5 (68)	1	6.6 (324)	4.9 (48)	1.7	0.7
Mother died	6 (531)	6.8 (33)	-0.8*	6.6 (324)	4.8 (9)	1.8	2.6

Number of observations are in brackets. Significance at 1% (\*\*\*), 5% (\*\*), 10% (\*)

1. Difference between differences in mean within the whole sample and the sample of non-tracked.

2. Orphans are defined as those who lose one or both of their parents before reaching 15 years old.

Regarding results on height, similarly to the paper by Beegle, De Weerd and Dercon (2008), we find that starting from a sample of non-orphaned children aged 6-15 years at baseline, those who subsequently lose a parent will, by adulthood, be on average smaller than those who do not, whatever the situation of orphans, and whatever the sample used; except for those who lose their mother in the whole sample. These differences are highly significant with a great order of magnitude for orphans who lose their father in both samples. Conversely, losing the mother tends to have no significant impact on the final height of children in both samples even though the difference tends to be favorable to the orphans.

More importantly, we observe that the difference in average height between non-orphans and orphans is bigger in the restricted sample of non-tracked compared to the whole sample. This is true for all types of orphan-hood status.

As we can see by considering both samples, differences between non-orphans and orphans do change a lot specifically for father orphan-hood. The differences between the average differences yielded by each sample given the type of orphan-hood are always positive and large. This difference mechanically comes from differences in level. In fact, average height of non-orphans tends to be higher in the restricted sample than in the whole sample. Conversely, average heights of orphans are lower in the restricted sample of non-tracked individuals than in the whole sample. This discrepancy certainly provides some first hints about the source of the differences observed.

Results regarding education are rather different from what we observed with health. In fact, almost none of the differences are significant in both samples. We only have one exception with mother orphan-hood in the whole sample, but the sign is unexpected. This sign is in line with what we observed in the case of height. These results are quite different from the ones of the reference paper by Beegle et.al (2008). In fact they are different in terms of significance, magnitude and sign. This is not surprising since we do not use the same sample and the same concept of orphan-hood as in the

reference article. Even though the results are not significant we can still notice that the impact of orphan-hood is higher in the restricted sample than in the whole sample and “of the good sign”, except for mother orphan-hood (as the sign of the differences is positive. More generally regarding the impact of orphan hood on schooling, there might be two possibilities. Either, the average differences computed are biased or they constitute definitely the causal impact of orphan-hood on the education level.

In the first case, we can expect to find significant results by introducing some omitted variables in an OLS regression to correct for this, expecting that the difference in the differences would be significant suggesting that attrition or differences in the baseline variables is affecting the impact of orphan-hood on education. In the second case, we should refrain from comparing two non significant differences since they are statistically nil and as a consequence equivalent.

Above all these results about education, it is important to point out one key particularity of the education data. In fact, we have observed<sup>1</sup> that nearly 70% of the educated individuals have their highest level of education as being the last primary school class. This overrepresentation of P7<sup>2</sup> grades individuals is observed in the restricted sample of non-tracked individuals as well. Indeed, almost all the children have gone to primary school regardless their gender. According to a paper by Porter Karen, “Tanzania is one of the few countries in sub-Saharan Africa to have achieved near gender equity in primary schools. The abolition of primary school fees in 1973 removed that impediment to schooling, and Education Act no. 25 of 1978 made enrollment and attendance of boys and girls in primary school compulsory. All villages in Tanzania have at least one primary school.” Hence, there is likelihood of non-significant difference between the average years of schooling of non-orphans and orphans within the restricted sample.

All these perspectives will be enlightened in the next section, where we are going to try to find out if orphans and non-orphans are in fact different with regard to their baseline characteristics within both samples. If orphans and non-orphans are different within both samples, then we would need to control for these characteristics in our estimation of mean height and mean schooling in 2004. Comparing then the estimates of both samples will be thus robust to these initial differences. If after doing this, the impact of orphan hood still differs across samples, it would mean that selection is on unobservables and that there is clearly a selection bias due to attrition.

#### **b. Making sure that we are comparing similar samples: running descriptive statistics**

Our first attempt to check for selection bias due to attrition is somehow useful but the validity of the previous results requires similar samples of orphans and non-orphans in 1991 within each sample (i.e. whole sample, and the sample without tracked individuals). Otherwise we cannot interpret the results as causal effects of orphanhood since we have not controlled for baseline characteristics.

In order to approach the intrinsic causal effect of orphan-hood on human capital dimensions, we would need both samples of orphan and non-orphan to be similar with respect to observables and non-observables baseline characteristics. Given that unobservables are out of our reach, we can nevertheless control for some observables baseline characteristics which are not similar for the two groups and correlated to the outcomes. We divide the relevant baseline characteristics into three groups. The first group includes variables that are not specific to an outcome of interest. It encompasses the age, the sex, and the proportion of individuals living with their parents. In fact,

---

<sup>1</sup> Please refer to appendix 1 for results.

<sup>2</sup> The final class in primary school.

younger children will tend to be smaller or have less years of schooling. Similarly, if sex dimorphism or discrimination holds, being male or female will affect the final height attainment or final years of schooling. In addition, the fact of living with a given parent or both can heavily affect the health and/or education of the children since parents are the first source of care. The second group is about variables related to height whereas the third group includes variables specific to the education attainment. The selection of these variables is constrained by the available information and there could be unobserved characteristics at play in the results. Nevertheless, we focus on observables characteristics with the following results provided by tables 2.a, 2.b and 2.c.

Generally speaking, the similarity between orphans and non-orphans in terms of their baseline characteristics depends on the sample used and baseline characteristics. In fact, regardless of the parent that died (i.e. alluding to any parent died), results from Table 2.a imply that non-orphans were older, more educated, and fewer live with their father compared to orphans. The fact that non-orphans are older probably results from our definition of orphan-hood which encompasses all individuals who lost one or both of their parents before the age of 15. Therefore, the orphans as measured are likely to be the younger children in 1991. Results from table 2.b suggest that parental orphans had the same profile as described previously. In addition, most of them were female, and smaller as opposed to non-orphans. These latter correlates are compatible with the sexual dimorphism in favor of male children. Still, the fact of being smaller may be due to their age since orphans were younger in baseline. Dealing with mother orphan-hood, we find from table 2.c that it is mostly random with respect to the observables baseline characteristics. However, some of the profiles of orphans in the previous cases still hold.

Since we observe some differences at baseline and since these differences are correlated to the orphan hood status and the outcomes in 2004, we run an OLS regression controlling for the variables. The final goal as mentioned is to correct for any bias in the estimates of orphan hood on final outcomes within each sample in order to have a robust a comparison of these estimates across samples.

Table 2.a Difference in baseline characteristics by future orphan status (any parents died before 15)

	Whole sample			Sample without tracking		
	Remain non-orphan to age 15 n=531	Loses one or both parents by age 15 n=118	Difference: Non-orphan vs Orphan	Remain non-orphan to age 15 n=324	Loses one or both parents by age 15 n=69	Difference: Non-orphan vs Orphan
Baseline characteristics (1991)						
Age (years)	11,24	9,25	1,99***	11,05	9,22	1,83**
Male (%)	37	33	4	51	15	36**
Living with father (%)	75	94	-19**	84	96	-12*
Living with mother (%)	81	60	21	92	57	35*
Height (cm)	132,3	126,2	6,1	131,4	123	8,4*
Household per capita health expenditures (Log Tanzania shillings)	0,62	0,06	0,56	0,17	0,02	0,15**
Schooling (years)	1,8	0,4	1,4***	1,5	0,3	1,2***
Household per capita schooling expenditures (Log Tanzania shillings)	4,5	2,9	1,6	3,5	0,7	2,8***

Standard deviations are in parenthesis. Significance at 1% (\*\*\*), 5% (\*\*), 10% (\*)

Table 2.b Difference in baseline characteristics by future orphan status (Father died before 15)

	Whole sample			Sample without tracking		
	Remain non-orphan to age 15 n= 531	Loses father by age 15 n=68	Difference: Non-orphan vs Orphan	Remain non-orphan to age 15 n=324	Loses father by age 15 n=48	Difference: Non-orphan vs Orphan
Baseline characteristics (1991)						
Age (years)	11,24	9,37	1,87**	11,05	9,38	1,67*
Male (%)	37	9	0,28***	51	5	46***
Living with father (%)	75	96	-21***	84	96	-12*
Living with mother (%)	81	52	29	92	50	42*
Height (cm)	132,3	124,1	8,2**	131,4	123,8	7,6
Household per capita health expenditures (Log Tanzania shillings)	0,62	0,05	0,57	0,17	0,02	0,15**
Schooling (years)	1,8	0,3	1,5***	1,5	0,3	1,2***
Household per capita schooling expenditures (Log Tanzania shillings)	4,5	1	3,5***	3,5	0,7	2,8***

Standard deviations are in parenthesis. Significance at 1% (\*\*\*), 5% (\*\*), 10% (\*)

Table 2.c Difference in baseline characteristics by future orphan status (Mother died before 15)

	Whole sample			Sample without tracking		
	Remain non-orphan to age 15 n=531	Loses mother by age 15 n=33	Difference: Non-orphan vs Orphan	Remain non-orphan to age 15 n=324	Loses mother by age 15 n=9	Difference: Non-orphan vs Orphan
Baseline characteristics (1991)						
Age (years)	11,24	9,4	1,84***	11,05	11,15	-0,1
Male (%)	37	62	-25	51	59	-8
Living with father (%)	75	92	-17**	84	94	-10
Living with mother (%)	81	65	16	92	99	-7**
Height (cm)	132,3	131,8	0,5	131,4	130,2	1,2
Household per capita health expenditures (Log Tanzania shillings)	0,62	0,07	0,55	0,17	0	0,17***
Schooling (years)	1,8	0,5	1,3***	1,5	1,8	-0,3
Household per capita schooling expenditures (Log Tanzania shillings)	4,5	6,4	-1,9**	3,5	4,9	-1,4

Standard deviations are in parenthesis. Significance at 1% (\*\*\*), 5% (\*\*), 10% (\*)

## B. Econometric estimation: Assessing selection and importance of tracking

### a. Improving our first attempt results to detect selection

The previous part has shown that orphans and non-orphans are not similar in terms of all observables determinants in 1991. Therefore we need to correct for an omitted variable bias by adding controls to the simple linear equation, in order to have a robust comparison across samples. The general equation to be estimated reads:  $Y_{04i} = \alpha + \beta orph_i + X_{91i}\gamma + \varepsilon_i$

Where  $Y_{04i}$  is the outcome in 2004,  $orph$  is a dummy indicating whether an individual is orphan or not.  $X_{91i}$  is a vector of controls at baseline.

However, before running the OLS estimation, we need first to ensure that these controls are not correlated with the selection process. Otherwise, the impact would still be biased. To this end, tables 3 and 4 report mean comparisons and OLS regressions results respectively.

First of all, the distribution of individuals given their orphan-hood status is independent from their migration status as shown by the last column of the first row in table 3. The p-value attached to the Chi-square statistics is higher than 10%. Therefore the decision to migrate does not depend upon individuals' orphan-hood status. This ensures that the point estimate of the principal equation is not biased by an eventual correlation between orphan-hood status and migration status. In addition, mean comparison between tracked and non-tracked of the explanatory variables used leads to the conclusion that they are not correlated with migration decision of individuals. As a consequence, there is no bias in controlling for these variables.

Given these facts we have come up with the following conclusions: regardless of the parent who died (i.e. alluding to any parent died), the impact of orphan-hood on height differs from the whole sample to the restricted one. The point estimates are significant and amount for 3.81 cm less for orphans in the whole sample whereas the orphans are predicted to be 6.01 cm smaller than non-orphans. Therefore, there is a positive selection bias in using the restrictive sample to estimate the impact of orphan-hood regardless of the parent who died. More specifically, the death of a father seems to have the most contradictory impact depending on the sample used: Non-significant in the whole sample, but significant with a magnitude of 5.13 cm in the non-tracked sample. Then sample selection not only affects the significance of the impact but also its magnitude. The direction of the bias is positive as well. Conversely, mother bereavement is non-significant in both samples suggesting that we would have come to the same conclusion about the impact of mother orphan-hood had there been no tracking. As a result, there is sample selection issue affecting the measurement of the impact of father orphan-hood.

Results with respect to education are still non-significant irrespective of the type of orphan-hood and the sample used. They suggest that sample selection is not an issue with the estimation of the impact of bereavement on education outcome. Though the magnitude of the point estimate differs from the whole sample to the restricted one, we would have come up to final conclusion that orphan-hood does not affect education outcome whether or not individuals were tracked. Still, our conclusion depends on the perspective at stake. Here we are considering a situation in which some individuals were not tracked. Actually, we could have found a significant impact of orphan-hood on education using the sample of tracked. Therefore, the fact that the whole sample and the sample of non-tracked provide non-significant impact does not imply that tracked and non-tracked are similar in the way they are affected by the orphan-hood. There might be some unobservable factors making the difference in the way education outcomes of these two populations are affected by bereavement. Though, orphan-hood seems not explaining education, we have strong evidence that initial health status affects final educational attainment. In fact, results in table 4 suggest that the impact of initial health status on final education is not affected by selection bias and it is robust to the type of orphan-hood in terms of magnitude and precision.

As a result, we can conclude that there is an issue of selection bias affecting the estimated impact of parental orphan-hood on human capital. Now, we turn the final phase which is to check the possibility to correct for this bias.

#### **b. Exploring tools to correct for selection bias**

From this point on we focus, without loss of generality, our analysis on the selection bias affecting the impact of father orphan-hood on final height in 2004. This choice stems from the fact the between-sample comparison of this impact yield the most striking result: non-significant in the whole sample whereas it is significant and large in the restricted sample. In addition, we suspect that results concerning mother orphan-hood may not be accurate because of the very fewer cases of maternal orphan-hood compared to parental one. Moreover, we would have chosen "any parent died" for this sample size reason but we think that using father orphan-hood is more concrete.

Turning now to the source of non-random selection, we suspect that selection is likely to be on unobservables. This is so because the probit estimation in appendix 2 suggests that there is no correlation between attrition and the observable variables: age, sex, height and years of schooling. It turns out that using the Heckman model under these conditions seems appropriate. But in our case we know from the results in table 3 that this is not straightforward since we found that individuals are equally likely to remain in the sample or to migrate whatever the age, sex, height and years of schooling of the individual. Then we need to find an instrument in order to correctly identify the parameter of orphan-hood variable.

We need to state this formally:

Indeed, the selection equation reads:  $A_i = Z_i\delta + v_i$ ; where  $Z_i$  is the vector of explanatory including the instrument. Using the fitted values from this regression to correct for selection, we end up with the following equation, which is the general form of a selection-corrected equation:

$$y = x\beta + \rho\sigma\lambda(z\delta) + \eta$$

Where  $\lambda(z\delta)$  is the inverse Mills ratio of the fitted values from the selection equation. The identification problem stems from the fact that  $Z$  may be correlated with  $x$ . Therefore an instrument is needed to overcome this problem.

The instrument we need must be correlated with the tracking status and the final height but not with father orphan-hood. It turns out that a valid candidate could be the variable indicating whether a child was living with their mother in 1991. In fact, a child living with their mother in 1991 is likely to remain in the sample unlike another who is not living with their mother. Living with the mother could be an indicator of less mobility since women are less mobile than men. Actually the later are more likely to be looking for job or more education so that they frequently migrate. However this is not the case with women. Even when the husband needs to migrate for job or higher education, the wife usually stays in the original household in order to take care of the children. Therefore, living with mother is correlated with the migration status. Results from last line of table 3 confirm this correlation.

Regarding the final height of children which is a proxy for their health status, children living with their mother in 1991 are likely to be healthier than those living without their mother. It is generally accepted that women care about the health status of their children more than men. Therefore, living with the mother is a positive advantage for those children in terms of their health status. Hence, there is a correlation between living with the mother in 1991 and the final height in 2004.

When it comes to the relationship between living with the mother and losing one's father, there is *a priori* no relationship between these two events. Empirically, results from table 2.b confirm the fact that there is no correlation between living with the mother and father orphan-hood.

As a result, the variable *living with the mother* can be used as an instrument to implement the Heckman procedure. Appendix 3 reports the result of the Heckman estimation. We can see that after trying to correct for the selection bias we do not have the same impact as in the whole sample. As a matter of fact we still have an impact which is similar both in terms of magnitude and precision with the one in the restricted sample. Therefore the Heckman procedure does not help to correct for the selection bias, though we find a valid instrument.

We must be however very careful about this conclusion. Indeed, one could think that there might be other observables determinants of migration. We will show however in the next part that if there are other observable determinants of migration in 1991, then the correlation is likely to be spurious.

Since we come up to the conclusion that Heckman procedure cannot be applied to correct for selection bias in this case, we suspect that there might be some unobservables determinants of migration at work. They are not only unobservable in 2004 but also in 1991. These variables could be some unobservable variables at baseline or some shocks occurring within the period between 1991 and 2004. For this latter category of unobservables, one rationale for their existence could be the time elapsed between baseline survey and the year of migration. Indeed, the decision to migrate several years after the baseline might not be linked at all with the initial characteristics of the individuals. Typically, a large share of individuals migrated after 1996 which is 5 years after the baseline survey. Hence characteristics in 1991 are not likely to affect the decision to migrate five

years later. In so far as we know that migration decision is not affected by orphan-hood status, then we expect to see some purely exogenous shocks occurring within the period from 1991 to 2004 triggering migration at some date.

If we had not obtained the sample with the tracked individuals, we would be in a lot of troubles. Hopefully, we are not in this case and we can unveil the nature of these shocks and the reasons of their existence.

Apart from the Heckman correction, we could also think of reweighting the restricted sample in 2004 such that it is representative of the original one. However, this procedure is purely statistical and does not account for any event that might be affecting those who left their original location. In addition, we need to identify identical individual between migrants and non-migrants in order to allocate the weight of the missing individual to the one who stay. This operation requires then some similarity between individuals based on the original weighting system. However, since we suspect that the problem at stake is related to unobservables, then we cannot correctly reallocate the weight of missing individuals so as to obtain a representative sample of the original one.

The purpose of the remaining section of the study is to unveil the source of the selection bias given that the observable baseline characteristics cannot explain the selection process. To wrap up at this stage, we come to the partial conclusion that sample selection would bias upward the impact of father orphan-hood on height if tracking did not place. We have also seen that Heckman correction is not possible. In order to conclude on whether tracking is important we need to deepen our analysis by looking for the reasons of migration as we actually have the sample of tracked individuals.

Table 3. Patterns of attrition due to migration

	Sample of Non-Tracked		Sample of Tracked (Migrants)		Difference between Non-Tracked and Tracked <sup>2</sup>
	Non-Orphans	Orphans <sup>1</sup>	Non-Orphans	Orphans	
Proportion of individuals <sup>3</sup>	324 (87.10)	207 (91.19)	48 (12.90)	20 (8.81)	2.34 (0.126)
<b>Age in 1991</b>	11.05	9.37	11.40	9.43	0.42
<b>Proportion of male</b>	0.51	0.05	0.25	0.65	-0.21
<b>Height (cm) in 1991</b>	131.47	123.85	133.15	130.01	2.42
<b>Years of schooling in 1991</b>	1.52	0.24	2.04	0.99	0.56
Height in 04	163.40	153.64	158.09	165.54	-4.40
Years of schooling in 04	6.61	4.88	5.56	7.90	-0.88
Living with Mother in 04 (%)	0.16	0.05	0.004	0.40	-0.14***

Significance at 1% (\*\*\*), 5% (\*\*), 10% (\*). **It applies for the last column only.**

The four variables in bold characters are those likely to predict migration decision.

1. Only those who lost their father

2. Difference between the averages in the non-tracked sample and the tracked sample.

3. In first 4 columns, proportion of individuals in a given sample is in parenthesis. Chi-square statistics of distribution independence in the last column and p-value in parenthesis.

Table 4. OLS estimates of the impact of orphan-hood on human capital

	Whole Sample						Non-Tracked Sample					
	Height			Years of schooling			Height			Years of schooling		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Any parent died	<b>-3.59**</b> (1.85)			-0.39 (1.00)			<b>-5.76**</b> (2.66)			-2.20 (1.53)		
Father died		<b>-3.80</b> (2.78)			-0.50 (1.30)			<b>-4.84*</b> (2.99)			-1.84 (1.62)	
Mother died			<b>-2.24</b> (1.88)			0.47 (0.81)			<b>-1.93</b> (1.84)			-1.91 (1.85)
Age in 1991	-2.27*** (0.78)	-2.26*** (0.81)	-2.36*** (0.79)	-0.91*** (0.23)	-0.92*** (0.23)	-0.89*** (0.23)	-1.36*** (0.44)	-1.39*** (0.44)	-1.52*** (0.42)	-0.97*** (0.28)	-0.98*** (0.28)	-0.95*** (0.29)
Sex (1 for male)	8.86*** (1.66)	8.89*** (1.80)	8.94*** (1.76)	0.31 (0.68)	0.38 (0.72)	0.32 (0.71)	10.87*** (1.68)	11.22*** (1.69)	11.35*** (1.74)	-0.27 (1.02)	-0.14 (1.05)	-0.16 (1.08)
Height (cm) in 1991	0.41** (0.19)	0.41** (0.20)	0.41** (0.19)	0.07*** (0.02)	0.07*** (0.02)	0.07*** (0.02)	0.18* (0.09)	0.17* (0.09)	0.17** (0.08)	0.07*** (0.02)	0.07*** (0.02)	0.07*** (0.02)
Years of schooling in 1991	0.45 (0.50)	0.47 (0.51)	0.51 (0.51)	0.98*** (0.22)	1.00*** (0.23)	0.98*** (0.23)	0.51 (0.40)	0.53 (0.40)	0.62 (0.39)	0.80*** (0.20)	0.81*** (0.20)	0.79*** (0.21)

Robust standard errors are in parenthesis. Significance at 1% (\*\*\*), 5% (\*\*), 10% (\*)

(1): Using any parent death dummy

(2): Using Father death dummy

(3): Using Mother death dummy

Impact of both parents died on height in the sample of non-tracked is significant and equals (-11.67)

### c. Unveiling the source of selection

We have seen that the observables baseline characteristics cannot predict the migration status of an individual. Then we think of the unobservables at the baseline or within the period, specifically some events occurring during the year of migration. For both categories of unobservables, we take advantage of the data provided by the tracking process in order to uncover the deep cause of migration. The fact that migrants have been observed through thanks to tracking allow us to know more about them in order to explore the elements that can be part of the unobservables. However, we will first confirm the failure of the Heckman procedure since we can now relate our observables to the reasons for migration.

Figure 5 suggests that almost 3 out of 4 migrants moved for marriage reasons. Others migrate by following their spouse or parents, and the most of the rest migrate either to look for a job or school. If these alternatives are truly the rationale behind migration decision, then we should expect that age, sex, height and education at baseline are correlated with the migration process.

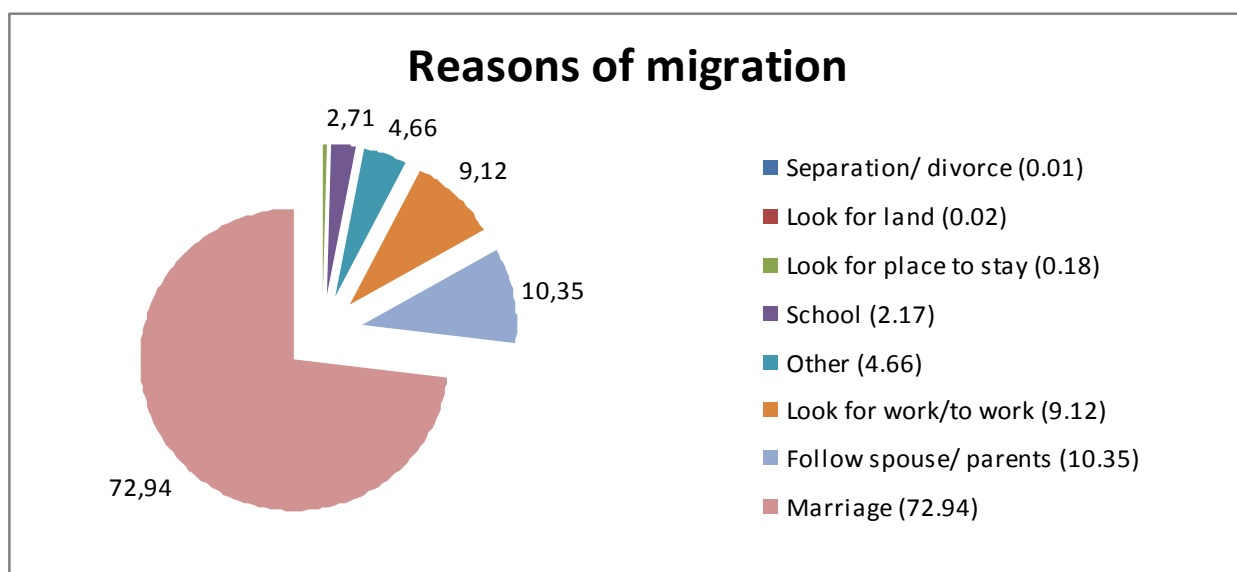
Indeed, since the largest part of migrants moved because of marriage, then the higher the age in 1991, the higher the probability to migrate. Individuals reaching the age of marriage in the initial cohort are likely to migrate in order to get married. In other words, migrants are expected to be older than non-migrants. Hence, there should be a positive relationship between migrants and non-migrants. However, figure in the second line, last column of table 3 suggests that there is no significant difference between the age of migrants and non-migrants.

On the same line with age, we would expect migrants to be taller than non-migrants since they are expected to be older. But there is no significant difference between the height of migrants and non-migrants. The same reasoning applies to sex because figures show that almost 100% of married migrants are women, hence we should expect sex to be correlated with migration decision.

Using schooling justifications, we might predict that the higher the level of schooling in 1991, the higher the probability to migrate before 2004. This is due to the scarcity of secondary school in Kagera and the fact that up to 81% of those who migrated for schooling reasons were in P5 level in 1991.

Above all, no significance difference has been found between migrants and non-migrants with respect to these observables. In addition, if our previous expectations were true, then we must observe rather uniform distribution of the migrants with respect to the year of migration for each reason. Surprisingly, figures in the table below show there are some sudden rise of migration flow during specific year. For instance, up to 85 percent of those who migrate because of marriage got married in 1999 or 2000. Similarly, most of individuals who followed their parents or spouse and those who are looking for job left in 2003. These results suggest that migration in the Kagera region is year specific and may result from some exogenous shocks. In the case of marriage, there may be an institutional change like a reduction in the cost of marriage license in 1999. There may also be some economic crisis resulting from drought or lack of resources in the Kagera's river in 2003 which trigger the migration of those looking for job and the other who follow their parents. The occurrence of these events makes the migration process independent from the baseline characteristics of individuals. Hence, had individuals been tracked, there is no way in using a Heckman selection equation to correct for the selection bias due to migration.

**Figure 5:** Reasons for migration



**Table 5:** Distribution of migrants according to the year and reasons for migration.

Year moved	Reason of migration			
	Marriage	Follow spouse/Parents	Look for Job	School
1993	0.00	0.52	0.00	1.71
1994	0.35	0.49	0.02	0.00
1995	5.87	0.00	0.09	3.21
1996	0.14	0.12	3.32	0.89
1997	0.68	0.22	1.49	0.53
<b>1998</b>	1.15	<b>17.79</b>	0.10	1.71
<b>1999</b>	<b>49.93</b>	0.13	<b>28.13</b>	0.28
<b>2000</b>	<b>35.71</b>	2.90	2.67	<b>80.60</b>
2001	0.36	1.06	1.26	0.66
2002	5.53	0.23	5.77	5.39
<b>2003</b>	0.28	<b>76.54</b>	<b>57.15</b>	5.01

## Conclusion

The purpose of this paper was to assess the importance of tracking individuals in long-term panel surveys. This is crucial from an analytical perspective because we can understand better the effect of attrition on the estimation of the impact of a given independent variable of interest on a given outcome.

In this paper, we assessed the importance of tracking individuals in long-term panel surveys focusing on the Kagera Health and Development Survey of Tanzania and by using the specific example of the estimation of the impact of orphan-hood on human development outcomes, such as height and education. The main question of this paper with regard to this empirical issue relating orphan hood to human development indicators was to know whether attrition in our sample was or not a source of selection bias. Thanks to the specific structure of our data we were able to answer to this question since we were able to compare the outcomes of the estimation using a sample without tracked individuals and a whole sample using tracked and non-tracked individuals, i.e. nearly the entire sampled population at baseline. By comparing these two samples under different econometric and descriptive applications we were able to assess the existence of a selection bias due to attrition.

The findings of this analysis deserve some attention. We found in this paper that there was a selection bias in our sample due to attrition. This selection bias was robust to the use of different techniques, descriptive and econometric. Moreover, we made sure in our analysis that the identification of this selection bias due to attrition was based on the fact the orphans and non-orphans were comparable at baseline within both samples. Had we failed to take care of this aspect, our identification of the selection bias due to attrition would be biased because of the endogeneity of the orphan hood variable.

This conclusion is however conditional on the issue at stake.

In the case of the relationship between orphan hood and height, we found that regardless of the orphan hood status, the impact of orphan-hood on height differs from the whole sample to the restricted one. The point estimates are significant and amount for 3.81 cm less for orphans in the whole sample whereas the orphans are predicted to be 6.01 cm smaller than non-orphans. Therefore, there is a positive selection bias in using the restrictive sample to estimate the impact of orphan-hood regardless of the orphan hood status.

More interestingly, the death of a father seems to have the most contradictory impact depending on the sample used: Non-significant in the whole sample, but significant with a magnitude of 5.13 cm in the non-tracked sample. Then sample selection not only affects the significance of the impact but also its magnitude. The direction of the bias is positive as well. Conversely, mother bereavement is non-significant in both samples suggesting that we would have come to the same conclusion about the impact of mother orphan-hood had there been no tracking. As a result, there is sample selection issue affecting the measurement of the impact of father orphan-hood.

In the case of the relationship between orphan hood and schooling results seem to suggest that we do not have selection bias due to attrition. Indeed, results are non-significant irrespective of the type of orphan-hood and the sample used.

Focusing on the sample of orphans of father (because the impact in this sample yielded the most impressive results), our analysis in this paper tried to go further. We tried indeed to unveil the source of non-random selection. In pursuing this goal, we found some evidence suggesting that in our sample, had we not tracked individuals, selection would be on unobservables. The migratory status of the individual was indeed orthogonal to the age, the sex, the height and the schooling of the non-

orphans in 1991. As it turns out, because we cannot predict the migratory status of individuals due to observables, it is difficult to apply this methodology. This finding is important because it suggests that any correction of selection bias should place emphasis on unobservable determinants. But this is unlikely since it is difficult to deal with the unobserved.

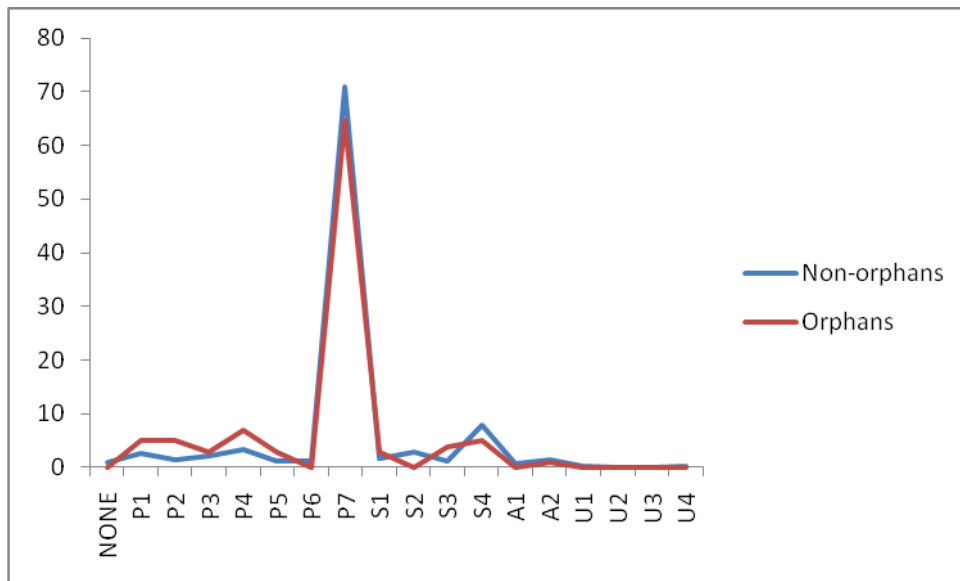
However as we have our data with tracked individuals we are able to know the true reasons of migration. In figure 5 we saw for example that almost 3 out of 4 migrants moved for marriage reasons. Others migrate by following their spouse or parents, and the most of the rest migrate either to look for a job or school. These results suggest that the reasons of migration are exogenous shocks that are moreover orthogonal to the observable characteristics in 1991 that could be predictors of migration.

The implications of this finding for methodologies that deal with selection bias are not negligible. Indeed, had we applied the Heckman methodology, even using other observable characteristics, we would have mistakenly conclude that migration was linked to events that occurred at baseline, or when the child became an orphan. And thus the correction would have been wrong since we know that the reasons individuals left the sample are unrelated to these baseline characteristics, and are closer to being exogenous shocks than endogenous in the equation. Taking as an alternative a weighting strategy again would have been problematic since 1991 baseline observable characteristics are not good predictors of the propensities of individuals to migrate. We are those unable to estimate unbiased propensities and to create counterfactuals.

Our paper concludes unfortunately that attrition is more seriously than expected even with the standard methodologies to deal with this issue. This means that we need to be more creative and come up with better solutions. It also means that we need a lot of resources to invest in surveys that track individuals.

## APPENDIX:

### Appendix 1: Graph on distribution of the level of education:



### Appendix 2: Probit estimation of selection equation with instrument

	Tracking status
Age	-0.057 (0.45)
Sex	0.617 (1.57)
Height in 1991	0.005 (0.27)
Highest grade in 1991	-0.060 (0.49)
Constant	-0.154 (0.09)
Observations	599
Robust z-statistics in parentheses	
* significant at 5%; ** significant at 1%	

### Appendix 3: Heckman estimation

VARIABLES	(1) height	(2) select	(3) athrho	(4) lnsigma
Father died	-5.102* (2.987)			
Age	-1.427*** (0.450)			
Sex	11.26*** (1.678)			
Height in 1991	0.179* (0.0925)			
Highest grade in 1991	0.584 (0.435)			
Living with mother in 1991		0.710* (0.440)		
Constant	147.8*** (9.385)	-0.642* (0.367)	0.271 (0.739)	1.637*** (0.154)
Observations	599	599	599	599

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1  
Robust standard errors in parentheses

### REFERENCES:

- Beegle, K., De Weerdt, J. and Dercon, S. 2008a. "Migration and Economic Mobility in Tanzania: Evidence from a Tracking Survey". Policy Research Working Paper, WPS 4798, World Bank, Washington DC.
- Beegle, K., De Weerdt, J. and Dercon, S. 2008b. Adult Mortality and Economic Growth in the Age of HIV/AIDS. *Economic Development and Cultural Change*, Vol. 56, No. 2: 299-326.
- Beegle, K., De Weerdt, J. and Dercon, S. 2008c. The Intergenerational impact of the African orphans crisis; a cohort study from HIV/AIDS affected area. *International Journal of Epidemiology*; 38:561-568.
- Evans, David and Edward Miguel. 2007. "Orphans and Schooling in Africa: A Longitudinal Analysis." *Demography* 44(1): 35-57.
- Fitzgerald, John, Peter Gottschalk and Robert Moffitt (1997). "An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics" <http://fmwww.bc.edu/ec-p/wp394.pdf>.
- Porter, Karen. 2003. "Gender and Education in Tanzanian Schools/Lessons from Kilimanjaro: Schooling, Community, and Gender in East Africa African". *Studies Review*.